ภาสพิชญ์ชูใจ : การเรียนรู้ร่วมกันสำหรับปัญหาการจำแนกข้อมูลไม่สมดุล (ENSEMBLE LEARNING FOR IMBALANCED DATA CLASSIFICATION PROBLEM)อาจารย์ที่ปรึกษา:รองศาสตราจารย์ ดร.นิตยาเกิดประสพ, 130 หน้า.

ข้อมูลไม่สมดุลเป็นข้อมูลที่สามารถพบเจอได้จริงในชีวิตประจำวัน เช่น ข้อมูลการวินิจฉัยโรคที่พบได้ยากทางด้านการแพทย์ เมื่อนำข้อมูลเหล่านี้มาใช้งานทางด้านการเรียนรู้ของเครื่องจักรและการทำเหมืองข้อมูลจะส่งผลกระทบต่อการเรียนรู้ของอัลกอริทึม เนื่องจากข้อมูลที่ใช้ในการเรียนรู้มีและเป็นกลุ่มที่ให้ความสนใจมีจำนวนข้อมูลที่น้อยมากเมื่อเทียบกับกลุ่มอื่น ๆ ที่เหลืออัลกอริทึมทางด้านการเรียนรู้ของเครื่องจักรนั้นสามารถทำงานได้ดีในกรณีที่ข้อมูลสมดุล สำหรับข้อมูลไม่สมดุลนั้นขอบเขตของการตัดสินใจของอัลกอริทึมการเรียนรู้ของเครื่องจักรนั้นจะมีความเอนเอียงไปทางกลุ่มข้อมูลส่วนมากส่งผลให้การจัดกลุ่มของข้อมูลส่วนน้อยมีแนวโน้มที่จะได้รับการจัดกลุ่มที่ผิดประเภทดังนั้นงานวิจัยนี้จึงนำเสนออัลกอริทึมใหม่ชื่อว่า EnsDTV (Ensemble_Learning_with_DecisionTree_Visualization) เพื่อแก้ปัญหาการจำแนกประเภทข้อมูลไม่สมดุลที่ข้อมูลอาจจะมีอัตราความไม่สมดุลของกลุ่มข้อมูลสูงและมีอัตราการซ้อนทับกันของกลุ่มข้อมูลที่แตกต่างกันด้วยการนำวิธีการเรียนรู้ร่วมกันแบบการใช้การตัดสินใจร่วมกันทั้งแบ็กกิง (Bagging)และบูสต์ติง (Boosting)มาทำการสร้างโมเดล ชดเชยการทำนายข้อมูลผิดกลุ่มด้วยวิธีการเรียนรู้แบบมีค่าใช้จ่าย (Cost-Sensitive Learning)ด้วยการนำค่าจากตารางค่าใช้จ่ายมาใช้ในขั้นตอนการเรียนรู้และปรับพารามิเตอร์ของการเรียนรู้ร่วมกัน ใช้โครงสร้างต้นไม้ตัดสินใจ (Decision Tree) เป็นเครื่องมือในการจำแนกข้อมูลพร้อมทั้งปรับลดจำนวนต้นไม้ตัดสินใจด้วยวิธีการสร้างมโนภาพหรือวิชวลไลเซชัน (Visualization) และการเตรียมข้อมูลให้เหมาะสมด้วยการลดการใช้พื้นที่ร่วมกันให้เบาบางลง ผลที่ได้ปรากฏว่า เมื่อนำวิธีการที่นำเสนอมาทำงานกับชุดข้อมูลที่มีการลดอัตราการซ้อนทับกันของกลุ่มข้อมูลที่เหมาะสมแล้วพบว่าสามารถนำมาใช้แก้ปัญหาการจำแนกประเภทข้อมูลไม่สมดุลที่มีอัตราความไม่สมดุลที่สูงและมีอัตราการซ้อนทับที่แตกต่างกันได้อย่างมีประสิทธิภาพ โดยเฉพาะโมเดลที่มีการเรียนรู้ร่วมกันด้วยการใช้การตัดสินใจร่วมกันแบบบูสต์ติงนั้นจะให้ประสิทธิภาพในการจำแนกประเภทข้อมูลกลุ่มส่วนน้อยได้ดีกว่าโมเดลที่มีการเรียนรู้ร่วมกันด้วยการใช้การตัดสินใจร่วมกันแบบแบ็กกิง ในขณะที่แบ็กกิงนั้นจะไม่สามารถทำงานได้เมื่อข้อมูลมีอัตราความไม่สมดุลที่สูงและมีอัตราการซ้อนทับที่ต่ำ

สาขาวิชาวิศวกรรมคอมพิวเตอร์ ลายมือชื่อนักศึกษา

ปีการศึกษา2557 ลายมือชื่ออาจารย์ที่ปรึกษา

PASAPITCH CHUJAI : ENSEMBLE LEARNING FOR IMBALANCED DATA CLASSIFICATION PROBLEM.THESIS ADVISOR: ASSOC. PROF.NITTAYA KERDPRASOP, Ph.D., 130 PP.

ENSEMBLE LEARNING/ DECISITION TREE / IMBALANCED DATA/ COST-SENSITIVE LEARNING/ VISUALIZATION

Imbalanced data area kind of data that can be found in real life, such as rare case in medical diagnosis. When used in machine learning and data mining,these data will affect the learning performance of algorithms.This is due to the amount of instances in the group of interest is much smaller than the other groups. In the field of machine learning, when data are balanced, a learning algorithm can be applied efficiently in terms ofoverall classification accuracy. For unbalanced data, the boundary of decision of most learning algorithms tend to bias toward the majority class andthe classification in the minority class will be misclassified. Therefore, we present a new technique called EnsDTV (Ensemble_Learning_with_DecisionTree_Visualization) for dealing with imbalanced classification problem: high imbalanced ratio and different overlapped ratio.To solve this problem,we apply the ensemble learningusing both bagging and boosting techniques to build models. We compensate themisclassification with cost sensitive learning and then use the value from cost matrix in the learning process to adjust the parameters of the ensemble learning. We adopt decision tree algorithm for data classification and reduce the number of decision trees by visualization.We prepared

optimal imbalanced dataset by reducing an overlapped region. The results showed that the proposed method work with datasets

thatreduction of the overlapped region then sends the EnsDTV method can solve the imbalanced data classification problem efficiently which high imbalance ratio and different overlapped ratio. Especially the ensemble learning using boosting techniques will enhance the classification minority classbetter than bagging. While the ensemble learning using bagging techniques cannot work in both case of high imbalance ratio and low overlapped ratio.

School of Computer Engineering Student's Signature

Academic Year2014 Advisor's Signature